Available online at www.sciencedirect.com

ScienceDirect

Journal of Hydro-environment Research xx (2014) 1–10

Journal of Hydro-environment Research

www.elsevier.com/locate/jher

Research paper

Copula-based modeling and stochastic simulation of seasonal intermittent streamflows for arid regions

Changsam Jeong [a], Taesam Lee [b,*]

[a] Dept. of Civil and Environmental Eng., Induk University, Wolgye 2-dong, Nowon-gu, Seoul 139-052, South Korea
[b] Dept. of Civil Engr., ERI, Gyeongsang National University, 501 Jinju-daero, Jinju, Gyeongnam 660-701, South Korea

Received 17 April 2013; revised 28 June 2014; accepted 30 June 2014

Abstract

Streamflow is often intermittent in arid and semi-arid regions. Stochastically simulated data play a key role in managing water resources with intermittent streamflows. The stochastic modeling of intermittent streamflow that incorporates the seasonality of key statistics is a difficult task. In the current study, the product model was tested to simulate the intermittent monthly streamflow by employing the periodic Markov chain (PMC) model for occurrence and the periodic gamma autoregressive (PGAR) and copula models for amount. The copula models were tested in a previous study for the simulation of yearly streamflow, resulting in successful replication of the key and operational statistics of historical data; however, the copula models have never been tested on a monthly time scale. The intermittent models were applied to the Colorado River system in the present study. A few drawbacks of the PGAR model were identified, such as significant underestimation of minimum values on an aggregated yearly time scale and restrictions of the parameter boundaries. Conversely, the copula models do not present such drawbacks but show feasible reproduction of key and operational statistics. We concluded that the copula models combined with the PMC model is a feasible method for the simulation of intermittent monthly streamflow time series.

Keywords: Copula; Drought; Intermittency; Periodic Markov chain; Seasonal streamflow; Stochastic simulation

1. Introduction

Synthetic data obtained from stochastic models play a key role in analyzing extreme events, such as droughts, and in evaluating alternative designs and operating rules of hydraulic structures, especially in arid or semi-arid regions (Lee and Salas, 2006; Salas, 1993; Salas and Abdelmohsen, 1993; Salas et al., 2006; Stedinger et al., 1983; Stedinger and Tasker, 1985). Many alternatives have been developed, which originated from a simple autoregressive model (Koutsoyiannis, 1994; Kwon et al., 2007; Lall, 1995; Lall et al., 1996; Lee and Ouarda, 2010; Lee and Salas, 2011; Lee et al., 2010; Ouarda et al., 1997; Salas and Abdelmohsen, 1993; Salas and Boes, 1980; Salas and Lee, 2010; Srinivas and Srinivasan, 2001, 2006; Stedinger and Vogel, 1984; Sveinsson et al., 2003).

A few characteristics remain difficult to reproduce, such as long-term persistency and intermittency. Long-term persistency (Lee and Ouarda, 2010; Sharma et al., 1998; Young and Holt, 2007) implies the auto-dependency structure of more than a monthly time scale (e.g., yearly). Intermittency (Koutsoyiannis, 2006) is identified by one or more zero values between non-zero values in a time series when values represent events or amounts of streamflows or rainfall. For example, in the Colorado River system, the monthly datasets of some tributary stations display intermittency.

Few effective models have been developed for generating monthly streamflow data with intermittency because of

* Corresponding author. Tel.: +82 55 772 1797; fax: +82 55 772 1799.
E-mail address: tae3lee@gnu.ac.kr (T. Lee).

http://dx.doi.org/10.1016/j.jher.2014.06.001
1570-6443/

structural difficulties inherent in modeling seasonality and intermittency. Chebaane et al. (1995) applied the product model, which comprises both amount and occurrence. The occurrence process has been described using the periodic discrete autoregressive (PDAR) model; the amount process has been described using the periodic autoregressive moving average (PARMA) model. The occurrence process must be modeled to preserve the statistical behavior of historical records. For example, the occurrence probability of a certain month in historical records should be reproduced by the simulated data.

We hypothesized that the physical process presented by the PARMA model is periodically stationary and that the marginal noise processes are normally distributed. Because monthly streamflow data are not normally distributed, the data needs to be transformed. The process of fitting the model to the transformed data and then back-transforming the data results creates bias toward the statistics simulated to describe the original space. Instead of this transformation, a model with a skewed distribution (i.e., gamma) was developed with preserving lag-1 serial correlation on a seasonal time series (Fernandez and Salas, 1986), namely the periodic gamma autoregressive (PGAR) model. However, the PGAR model requires a very complicated procedure to simulate the sequences, and its parameter space is limited. Meanwhile, Lee and Salas (2011) developed the copula-based stochastic model and applied it to a yearly time series. The results showed that the copula-based stochastic simulation model easily captures the key statistics of historical data and does not require any transformation. Furthermore, a number of applications for copula have been popularly reported in literature, including drought analysis (Shiau et al., 2007; Song and Singh, 2010; Wong et al., 2010) and flood frequency (Favre et al., 2004; Kao and Govindaraju, 2008; Zhang and Singh, 2006).

Therefore, we applied the copula-based model to monthly time scale data. In the current study, two models for the amount process (the PGAR and copula models) and one model for the occurrence process (the PDAR model) were tested to simulate intermittent monthly streamflow. The pros and cons of the models were inspected, and the model was applied to the Colorado River system.

2. Proposed modeling approach

To model an intermittent streamflow time series, the product of occurrence and amount is denoted as follows:

$$Y_{\nu,\tau} = X_{\nu,\tau} Z_{\nu,\tau} \tag{1}$$

where $\nu = 1,2,\ldots,N$ and $\tau = 1,2,\ldots,\omega$, representing years and seasons, respectively; N and ω are the numbers of years and seasons, respectively. Note that when the data time scale is monthly, $\omega = 12$. $X_{\nu,\tau}$ denotes the binary (0 or 1) occurrence process; $Z_{\nu,\tau}$ denotes the amount process; and $Y_{\nu,\tau}$ is the product of the two processes. The traditional PDAR(1) model was applied to simulate the occurrence process, $X_{\nu,\tau}$. The PGAR model was applied to simulate the amount process. The Bivariate Normal Copula with Gamma marginal distribution (BNCG) was also used to simulate the amount process.

2.1. Occurrence process

2.1.1. Order-1 periodic discrete autoregressive (PDAR(1)) model

The PDAR(1) is defined as follows:

$$X_{\nu,\tau} = V_{\nu,\tau} X_{\nu,\tau-1} + (1 - V_{\nu,\tau}) W_{\nu,\tau} \tag{2}$$

where $X_{\nu,\tau}$ is a periodic dependent Bernoulli process; and $W_{\nu,\tau}$ and $V_{\nu,\tau}$ are independent Bernoulli processes with probabilities $P[V_{\nu,\tau} = 1] = \gamma_\tau$ and $P[W_{\nu,\tau} = 1] = \delta_\tau$, respectively. Chebaane et al. (1995) showed that the PDAR(1) model is equivalent to a periodic Markov chain (PMC) model, in which the elements of the transition probability matrix vary by season as

$$p_\tau(i,j) = P[X_{\nu,\tau} = j | X_{\nu,\tau-1} = i] \tag{3}$$

where $i, j = 0$ or 1.

The transition probability matrix is expressed as a function of the parameters of PDAR(1) such that

$$\begin{bmatrix} p_\tau(0,0) & p_\tau(0,1) \\ p_\tau(1,0) & p_\tau(1,1) \end{bmatrix} = \begin{bmatrix} \gamma_\tau + (1-\gamma_\tau)(1-\delta_\tau) & (1-\gamma_\tau)\delta_\tau \\ (1-\gamma_\tau)(1-\delta_\tau) & \gamma_\tau + (1-\gamma_\tau)\delta_\tau \end{bmatrix} \tag{4}$$

The limiting distribution $P[X_{\nu,\tau} = 1] = \mu_\tau$ is given by

$$\mu_\tau = \gamma_\tau \mu_{\tau-1} + (1 - \gamma_\tau)\delta_\tau \tag{5}$$

The parameters of PMC are estimated by

$$\widehat{p}_\tau(i,j) = \frac{n_\tau(i,j)}{n_\tau(i)} \tag{6}$$

where $n_\tau(i,j)$ is the number of times that the variable $X_{\nu,\tau}$ in state i at time $\tau - 1$ passes to state j during the period τ; and $n_\tau(i) = n_\tau(i,0) + n_\tau(i,1)$ is the number of times that $X_{\nu,\tau}$ is in state i at period τ. γ_τ and δ_τ are easily estimated from the relationship outlined in Eqs. (4) and (6).

2.2. Amount process

Two models were applied to simulate the amount process of the product model in the current study: (1) PGAR and (2) bivariate normal copula. These models are not based on Gaussian marginal distributions. Instead, PGAR assumes that the marginal distribution is gamma, while the bivariate normal copula model can have any feasible distribution for a marginal distribution. In practical applications, a gamma marginal distribution is applied for the bivariate normal copula even if there is no prior limitation for selecting marginal distribution because (1) the gamma distribution is one of the most frequently selected distribution types for hydrological frequency analysis; (2) it is comparable to the PGAR model for the marginal distribution; and (3) it is representative of the

Please cite this article in press as: Jeong, C., Lee, T., Copula-based modeling and stochastic simulation of seasonal intermittent streamflows for arid regions, Journal of Hydro-environment Research (2014), http://dx.doi.org/10.1016/j.jher.2014.06.001

positively skewed distributions that are typical in historical streamflow data.

For periodic models, Fourier series analysis may be used to model the periodic patterns of certain statistics. Fourier series analysis has been a popular method for use in stochastic models in which the time scale is relatively small, such as weeks and days. Katz and Parlange (1995) used this procedure to capture the diurnal cycle in hourly rainfall data. In the present study, the Fourier series was applied to smooth out periodic patterns, especially the skewness and lag-1 correlation, which have high degrees of uncertainty in parameter estimations.

2.2.1. Periodic gamma autoregressive (PGAR)

The gamma distribution is denoted as

$$f_z = \frac{1}{\alpha\Gamma(\beta)}\left(\frac{z-\lambda}{\alpha}\right)^{\beta-1}\exp\left(-\frac{z-\lambda}{\alpha}\right) \tag{7}$$

where $\alpha > 0$, $\beta > 0$ and $z > \lambda$. If $\lambda = 0$, then Eq. (7) presents a two-parameter gamma distribution represented as $Z \sim$ gamma(α, β and λ) or gamma(α, β). The relationships between the distribution parameters and the key statistics are

$$\beta = \left(\frac{\mu}{\sigma}\right)^2 \text{ and } \alpha = \frac{\mu}{\beta} \text{ for 2-gamma distribution} \tag{8}$$

$$\beta = \left(\frac{2}{g}\right)^2,\ \alpha = \frac{\sigma g}{2},\ \text{and}$$
$$\lambda = \mu - \alpha\beta \text{ for 3-gamma distribution} \tag{9}$$

where μ, σ and g represent the mean, standard deviation and skewness, respectively. PGAR was developed by Fernandez and Salas (1986) and is defined as

$$Z_{\nu,\tau} = \phi_\tau Z_{\nu,\tau-1} + Z_{\nu,\tau-1}^{\delta_\tau} E_{\nu,\tau} \tag{10}$$

where $Z_{\nu,\tau}$ is a continuous positive variable whose marginal distribution is gamma; ϕ_τ and δ_τ are seasonal autoregressive coefficients; and $E_{\nu,\tau}$ is the noise process. Note that for a three parameter case, $Z_{\nu,\tau}$ should be subtracted first by λ_τ and then applied to model Eq. (10).

The model parameters ϕ_τ and δ_τ, and the noise variable $E_{\nu,\tau}$ are given by

$$\begin{cases} \phi_\tau = 0 & \beta_\tau < \beta_{\tau-1} \\ \phi_\tau = \rho_{1,\tau}\dfrac{\alpha_\tau}{\alpha_{\tau-1}}\left(\dfrac{\beta_\tau}{\beta_{\tau-1}}\right)^{1/2} & \beta_\tau \geq \beta_{\tau-1} \end{cases} \tag{11}$$

$$\begin{cases} \delta_\tau = \rho_{1,\tau}\left(\dfrac{\beta_{\tau-1}}{\beta_\tau}\right)^{1/2} & \beta_\tau < \beta_{\tau-1} \\ \delta_\tau = 0 & \beta_\tau \geq \beta_{\tau-1} \end{cases} \tag{12}$$

and

$$\begin{cases} E_\tau = W_{\nu,\tau} & \beta_\tau < \beta_{\tau-1} \\ E_\tau = \varepsilon_{\nu,\tau} & \beta_\tau \geq \beta_{\tau-1} \end{cases} \tag{13}$$

where $\rho_{1,\tau}(z)$ is the lag-1 autocorrelation coefficient at period τ. Additionally, $W_{\nu,\tau}$ and $\varepsilon_{\nu,\tau}$ are defined in terms of α_τ, β_τ and $\rho_{1,\tau}$ so that $z_{\nu,\tau}$ has a lag-1 autoregressive dependence structure with periodic gamma marginal distribution. $W_{\nu,\tau}$ and $\varepsilon_{\nu,\tau}$ are simulated as follows:

(a) Generation of $\varepsilon_{\nu,\tau}$

$$\varepsilon_\tau = \varepsilon_\tau(0) + \varepsilon_\tau(1) \tag{14}$$

$$\varepsilon_\tau(0) \sim \text{gamma}(\alpha_\tau, \beta_\tau - \beta_{\tau-1}) \tag{15}$$

$$\begin{cases} \varepsilon_\tau(1) = 0 & M = 0 \\ \varepsilon_\tau(1) = \displaystyle\sum_{m=1}^{M} Y_m \left(\phi_\tau \frac{\alpha_\tau}{\alpha_{\tau-1}}\right)^{U_m} & M > 0 \end{cases} \tag{16}$$

where M is a Poisson random variable with the expected value $E(M) = -\beta_{\tau-1}\ln(\phi_\tau(\alpha_\tau/\alpha_{\tau-1}))$, $[U_m]_{m\in\{1,\ldots,M\}} \sim$ Unif(0,1) and $[Y_m]_{m\in\{1,\ldots,M\}} \sim$ Exp(α_τ). Here Exp(α) represents the random variate of exponential distribution with shape parameter (α).

(b) Generation of $W_{\nu,\tau}$

$$W_\tau = \frac{\alpha_\tau}{\alpha_{\tau-1}^{\delta_\tau}} s_\tau e_\tau^{\delta_\tau} \tag{17}$$

$$s_\tau \sim \text{Beta}(\beta_\tau, \beta_{\tau-1} - \beta_\tau) \tag{18}$$

$$e_\tau \sim \text{Exp}\left(V'_\tau + \lim_{K\to\infty}\sum_{k=1}^{K} V_{\tau,k}\right) \tag{19}$$

$$V'_\tau = \begin{cases} -\vartheta(1-\delta_\tau) & \text{w.p. } \delta_\tau \\ -\vartheta(1-\delta_\tau) - \text{Exp}(\beta_\tau) & \text{w.p. } 1-\delta_\tau \end{cases} \tag{20}$$

where ϑ is Euler's constant (=0.577216…)

$$V_{\tau,k} = \begin{cases} (1-\delta_\tau)/k - 0 & \text{w.p. } \delta_\tau \\ (1-\delta_\tau)/k - \text{Exp}(\beta_\tau + k) & \text{w.p. } 1-\delta_\tau \end{cases} \tag{21}$$

2.2.2. Copula-based time series modeling

A copula is the joint cumulative distribution function of a random vector with marginals that are uniform (Joe, 1997). The copula contains all of the information related to the dependence structure of its components. The copula concept makes it easier to formulate multivariate models compared to other complex and limited multivariate models. Therefore, in the current study, the ability of copulas to conveniently describe the dependence structure was employed to model the time dependence structure.

A yearly simulation model using the copula was described by Lee and Salas (2011). In the present study, the periodic copula model combined with the occurrence model was tested. Among a number of copula models, the bivariate normal (BVN) copula was employed to model periodic streamflow data at first. In the following section, the other copula functions were applied to each month. One copula model from the copula candidates as shown in Table 1 that shows the highest likelihood value was selected.

The mathematical description of the time series model with BVN copula is as follows.

Please cite this article in press as: Jeong, C., Lee, T., Copula-based modeling and stochastic simulation of seasonal intermittent streamflows for arid regions, Journal of Hydro-environment Research (2014), http://dx.doi.org/10.1016/j.jher.2014.06.001

Table 1
Commonly used copula models.

Type	$C(u,v)$	$C_{v\|u}$	Range
Frank	$-\frac{1}{\theta}\ln\left[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right]$	$-\frac{1}{\theta}\log\left[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{(e^{-\theta}-1)}\right]$	$\theta \neq 0$
Clayton	$(u^{-\theta}+v^{-\theta}-1)^{-\frac{1}{\theta}}$	$(u^{-\theta}+v^{-\theta}-1)u^{-\theta-1}$	$\theta \geq 0$
Gumbel	$\exp[-\{(-\log u)^{\theta}+(-\log v)^{\theta}\}^{1/\theta}]$	$\exp(-A^{1/\theta})A^{-1+1/\theta}(-\log u)^{\theta-1}u^{-1}$, $A=(-\log u)^{\theta}+(-\log v)^{\theta}$	$\theta \geq 1$
Bivariate normal	$\boldsymbol{\Phi}_{\theta}(\Phi^{-1}(u),\Phi^{-1}(v))$	$\Phi_{v\|u}$	$-1 \leq \theta \leq 1$

Note: $\Phi(x)=\int_{-\infty}^{x}(1/\sqrt{2\pi})\exp(-z^2)\mathrm{d}z$, $\boldsymbol{\Phi}_{\theta}$ is bivariate normal distribution with the correlation parameter θ, and $\Phi_{v|u}$ is the conditional normal distribution.

Assume that a continuous random variable $Y_{\nu,\tau}$ has a periodic Markov process such that,

$$F(Y_{\nu,\tau}|Y_{\nu,\tau-1},\ldots,Y_{1,1}) = F(Y_{\nu,\tau}|Y_{\nu,\tau-1}) = P\left(Y_{\nu,\tau}\leq y_{\nu,\tau}\middle|Y_{\nu,\tau-1}=y_{\nu,\tau-1}\right) \quad (22)$$

This implies that its past behavior does not influence the probability of any particular current behavior of the process except for the most recent previous condition. This Markov model of order 1 can be described using the copula. Let $F_{12}(y_{\nu,\tau},y_{\nu,\tau-1})=C(F(y_{\nu,\tau}),F(y_{\nu,\tau-1}))$ be a bivariate distribution with any univariate marginal distribution for each season, F_{Y_τ}.

The conditional distribution of the copula is denoted as

$$C(v|u)=\frac{\partial C(u,v)}{\partial u} \quad (23)$$

Then, the transition distribution of Y_τ, which is the periodic Markov model of order 1, is

$$F(Y_{\nu,\tau}|Y_{\nu,\tau-1}) = C\left(F_{Y_\tau}(y_{\nu,\tau})\middle|F_{Y_{\tau-1}}(y_{\nu,\tau-1})\right) \quad (24)$$

The representative copula models and the conditional distributions are presented in Table 1.

Once the BVN copula is selected for each season, the order 1 periodic Markov process (Eq. (23)) is presented (Chen and Fan, 2006) as follows:

(1) Fit a distribution model appropriate to the observed data for each season F_{Y_τ}, where $\tau=1,\ldots,\omega$ and ω is the number of seasons. Denote the cumulative density function of the historical data as $\widehat{F}_{Y_{\nu,\tau}}$
(2) Inverse the cumulative distribution function (cdf) of historical data into normal variate and denote it as $N_{\nu,\tau}$, i.e., $N_{\nu,\tau}=\Phi^{-1}(\widehat{F}_{Y_{\nu,\tau}})$, where $\Phi(N)=\int_{-\infty}^{N}(1/\sqrt{2\pi})\exp(-t^2)\mathrm{d}t$
(3) Generate $N_{\nu,\tau}$ as
$$N_{\nu,\tau}=\phi_\tau N_{\nu,\tau-1}+\varepsilon_{\nu,\tau} \quad (25)$$
where $\varepsilon_{\nu,\tau}$ is an uncorrelated normal variable with mean zero and variance $\sigma_\tau^2(\varepsilon)=1-\phi_\tau^2$, and ϕ_τ is the seasonal autoregressive coefficient of $N_{\nu,\tau}$ variable.
(4) Reinstate $N_{\nu,\tau}$ into the real domain by $Y_{\nu,\tau}=F_{Y_\tau}^{-1}(\Phi(N_{\nu,\tau}))$.

The following two sets of parameters are required: the parameters of the marginal distribution F_{Y_τ} and the parameters related to the $N_{\nu,\tau}$ process. The parameters for the marginal distribution can be estimated using any of the following estimation procedures: method of moments, maximum likelihood, L-moments and probability weighted moments from frequency analysis (Kottegoda and Rosso, 2008; Salas et al., 2008). The parameter of the copula model for the time series $N_{\nu,\tau}$ process can also be estimated using the method of moments or least square methods. An alternative method of estimating the inference function for margins (IFM) has been suggested in the copula literature (Joe, 1997; Nelsen, 1999).

For the other copula, the application is straightforward with the fitted marginal distribution $(\widehat{F}_{Y_{\nu,\tau}})$ from step (1) above as follows:

(a) Simulate the initial value from the marginal distribution $F_{Y_{1,1}}$, gamma distribution.
(b) Simulate the subsequent values $F_{Y_{\nu,\tau}}$ with the conditional distribution of Eqs. (23) and (24). The detailed equations are shown in Table 1.
(c) Inverse-transform the simulated values $F_{Y_{\nu,\tau}}^{-1}$ to obtain $Y_{\nu,\tau}$.

2.3. Parameter estimation and Fourier transformation

Model parameters must be estimated for the PGAR and BVN copula with gamma marginal parametric models. The method of moments was used for the parameter estimation in the current study for the marginal distributions. The method of moments estimates the model parameters according to the relationships between the statistical moments of observed data and the moments of the statistical models. The copula parameters in Table 1 were estimated from the inference functions for the margins (IFM) method (Joe, 1997).

The model applied in the current study is the product model shown in Eq. (1). The parameters of the amount process (i.e., Z variable in Eq. (1)) are estimated only using non-zero data. Therefore, the zero values in calculating the statistics must be excluded as:

$$\widehat{\mu}_\tau=\frac{1}{N_\tau^*}\sum_{\nu=1}^{N}z_{\nu,\tau}I_{z_{\nu,\tau}\neq 0} \quad (26)$$

$$\widehat{\sigma}_\tau^2=\frac{1}{N_\tau^*}\sum_{\nu=1}^{N}(z_{\nu,\tau}-\widehat{\mu}_\tau)^2I_{z_{\nu,\tau}\neq 0} \quad (27)$$

$$\widehat{g}_\tau=\frac{1/N_\tau^*\sum_{\nu=1}^{N}(z_{\nu,\tau}-\widehat{\mu}_\tau)^3I_{z_{\nu,\tau}\neq 0}}{\widehat{\sigma}_\tau^3} \quad (28)$$

Please cite this article in press as: Jeong, C., Lee, T., Copula-based modeling and stochastic simulation of seasonal intermittent streamflows for arid regions, Journal of Hydro-environment Research (2014), http://dx.doi.org/10.1016/j.jher.2014.06.001

where $I_{z_{\nu,\tau}\neq 0}$ is the indicator function such that $I_{z_{\nu,\tau}\neq 0}=1$ if $z_{\nu,\tau}\neq 0$, otherwise $I_{z_{\nu,\tau}\neq 0}=0, N_\tau^*=\sum_{\nu=1}^{N} I_{z_{\nu,\tau}\neq 0}$ and $\widehat{\mu}_\tau$, $\widehat{\sigma}_\tau^2$, $\widehat{g}_\tau$ are the sample mean, variance and skewness, respectively, for the period τ. The autocorrelation function $\rho_{k,\tau}$ is estimated as

$$\widehat{\rho}_{k,\tau}=\frac{\frac{1}{N_\tau^{**}}\sum_{\nu=1}^{N}(z_{\nu,\tau}-\widehat{\mu}_\tau)(z_{\nu,\tau-k}-\widehat{\mu}_{\tau-k})I_{z_{\nu,\tau}z_{\nu,\tau-k}\neq 0}}{\widehat{\sigma}_\tau\widehat{\sigma}_{\tau-k}} \tag{29}$$

where k is the time-lag; and $N_\tau^{**}=\sum_{\nu=1}^{N} I_{z_{\nu,\tau}z_{\nu,\tau-k}\neq 0}$.

The high degree of uncertainty, especially in the high-order moments (e.g., skewness and autocorrelation), may be assumed to eradicate the smooth seasonality. Therefore, Fourier transformation is applied to statistics with high degrees of uncertainty. In other words, it is assumed that parameters vary smoothly over the course of a month, a pattern that is referred to as seasonality. Furthermore, the parameter space of the PGAR model is highly limited. It is unable to apply the PGAR model when the estimated sample moments of the dataset fall outside of this parameter space. Fourier transformation can be applied to avoid this. Fourier transformation modifies these parameters to allow them to be located inside the parameter space (Fernandez and Salas, 1986). The Fourier transformation procedure is defined as

$$u_\tau^*=\overline{u}+\sum_{m=1}^{h} A_m\cos(2\pi m\tau/\omega)+B_m\sin(2\pi m\tau/\omega) \tag{30}$$

where u_τ^* is the Fourier transformed statistic from u_τ; $\overline{u}$ is the mean value of u_τ; and

$$\begin{aligned} A_m&=\frac{2}{\omega}\sum_{\tau=1}^{\omega} u_\tau\cos(2\pi m\tau/\omega)\ \text{ and }\ B_m\\ &=\frac{2}{\omega}\sum_{\tau=1}^{\omega} u_\tau\sin(2\pi m\tau/\omega);\quad m=1,\ldots,h \end{aligned} \tag{31}$$

When ω is even, the last coefficients are given by

$$A_h=\frac{1}{\omega}\sum_{\tau=1}^{\omega} u_\tau\cos(2\pi h\tau/\omega);\ B_h=0 \tag{32}$$

where h is the degree of smoothing. Note that if h gets smaller, the Fourier transformed statistics are smoothed more and vice versa.

3. Data description

The proposed models were applied to the Colorado River system. The Colorado River is a major river system in the western United States (US), and the Bureau of Reclamation uses 29 gauging sites within the system for long-term planning studies. The first 20 stations are categorized as Upper Colorado River stations, while the rest are Lower Colorado River stations. The Colorado River flows through arid and semi-arid regions of western states, such as Nevada and Arizona. The highly arid climate induces intermittent streamflow in certain months.

Among the stations, the monthly streamflow data of two stations, the Little Colorado River near Cameron, Arizona (USGS station number: AF09402000) and the Bill Williams River below Davis Dam located between Arizona and Nevada (USGS station number: AF09426000), show intermittency. The monthly data of the Little Colorado River station is intermittent every month, while the Bill Williams River site shows intermittency during only some months. Therefore, in the current study, the Little Colorado River site was used to validate model performance.

4. Application methodology

Two hundred samples were simulated using each model for a time period of the same length as the historical data (101 years). Two different models for the amount process were tested, PGAR and BNCG (Bivariate Normal Copula with Gamma marginal distribution), while the occurrence process was modeled with PMC. The gamma distribution was selected as the marginal distribution of the streamflow and is comparable to PGAR, which has a marginal that conforms to the gamma distribution.

Basic statistics such as the mean, standard deviation, skewness, maximum, minimum and lag-1 correlations in the seasonal and yearly time scales were estimated using the historical and simulated data to verify model performance. Using the water demand level as the mean, operational statistics such as maximum deficit (surplus) lengths and maximum deficit (surplus) amounts were also compared. Along with the water demand level, the storage capacity was also estimated using the sequent-peak algorithm (Loucks et al., 1981).

Furthermore, the nonparametric densities were estimated using the historical and simulated data. Because streamflow data cannot have values less than zero, the estimated densities are expected to be left-bounded. Subsequently, a boundary biweight kernel (Simonoff, 1996) was employed to estimate the density for the bounded-type distribution. If the density of the continuous variable $Y\in\{0,\infty\}$ needs to be estimated and the selected basic kernel, $K(\cdot)$, is a biweight kernel with a range of $[-1, 1]$, then the boundary kernel is adjusted over the range $0\le y<h$ with the equation

$$B(y)=\frac{[a_2(p)-a_1(p)]K(y)}{a_2(p)a_0(p)-a_1(p)} \tag{33}$$

where $a_l(p)=\int_{-1}^{p} u^l K(u)\mathrm{d}u$, $p=y/h$ and h is the smoothing parameter (bandwidth). The oversmoothing approach (Simonoff, 1996) was employed to select the bandwidth because the densities were used to compare the overall preservation of the historical density rather than to reproduce the small bumps.

Boxplots were employed to display the variability of the simulated statistics compared to that of the historical statistics. The end lines of the box (inter-quartile range) indicate the 25th

Please cite this article in press as: Jeong, C., Lee, T., Copula-based modeling and stochastic simulation of seasonal intermittent streamflows for arid regions, Journal of Hydro-environment Research (2014), http://dx.doi.org/10.1016/j.jher.2014.06.001

and 75th percentiles, while the cross lines above and below the box on the whisker denote the 90th and 10th percentile and the maximum and minimum, respectively. The dotted line connecting the 'x' marks represents monthly historical statistics, and the circles represent historical statistics in yearly time scale results.

5. Results of simulation model comparisons

In Figs. 1 and 2, the key statistics of the historical data and the simulated data from the PGAR and BNCG, respectively, are presented. Both models accurately reproduced the means and standard deviations of the historical data. The magnitude of skewness was negatively related to the mean; the skewness was lower during high flow months (March and April) than during low flow months (July and June) and vice versa. Both models slightly underestimated skewness during low flow months and overestimated skewness during high flow months. The lag-1 correlation of the historical data was reproduced accurately by all of the models as shown in Figs. 1 and 2, excluding a slight bias in some months. Extreme historical statistics were also reproduced accurately by both models. The minimum flow of every month was zero because of intermittency.

Fig. 3 illustrates the mean of the skewness (top panel) and lag-1 correlation (bottom panel) of the data simulated by the PGAR model for the amount process only (i.e., excluding zero values). As mentioned in the Methodology section, Fourier transformed skewness and lag-1 correlation must be employed to estimate the parameters of the PGAR model so that they will fall inside the parameter space of the PGAR model. Therefore, the PGAR model reproduces the Fourier transformed statistics instead of the real historical statistics.

The month-to-month relationships of the PGAR and BNCG models are shown in Figs 4 and 5, respectively. The gray circles represent the simulated data, while the inverted triangles show the historical values. The scatter plot of the PGAR model in Fig. 4 shows peculiar behavior, as the simulated values were not placed in the lower region of the straight line except for the zero values, which were simulated from the occurrence process (X). The slope of the straight line corresponds to the lag-1 serial correlation of month 6 ($2/3 \approx 0.67$). This phenomenon occurs because of the characteristics of the PGAR model shown in Eq. (10). Because the second term on the right side of Eq. (10) is always greater than zero, the simulated value is always greater than the lag-1 month-to-month correlation times of the previous month's value ($\phi_\tau Z_{v,\tau-1}$). Conversely, the BNCG model reproduces the historical relationships reasonably well, as shown by the simulated data in Fig. 5. Note that the simulated zero values for month 6 are located at the x-axis of Fig. 5, while no historical zero values were found. This discrepancy between the simulated and historical data is caused by the assumption of independence between the occurrence and amount process shown in Eq. (1). The same characteristic can be observed in Fig. 4 for the PGAR model.

Yearly key statistics are shown in Figs. 6 and 7 using the summed monthly simulated data to yearly data from the PGAR and BNCG models, respectively. Both models accurately reproduced the historical yearly statistics (circles), excluding

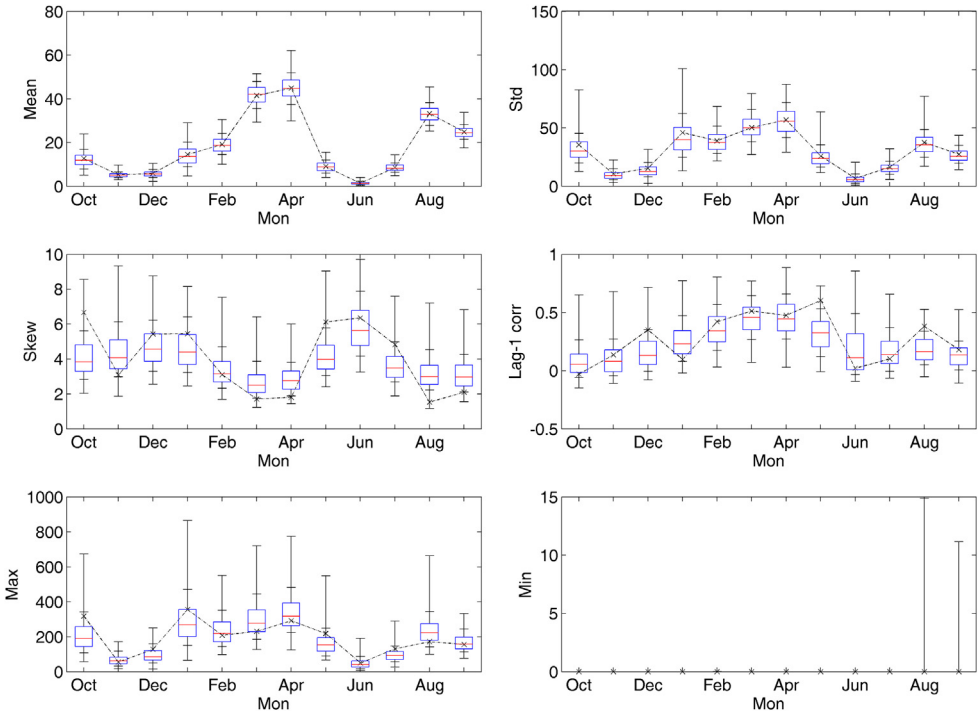

Fig. 1. Box plots of basic statistics for Colorado River monthly flows (million cubic meters, 10^6 m^3) obtained from data simulated using the PGAR model and from historical data (dotted line).

Please cite this article in press as: Jeong, C., Lee, T., Copula-based modeling and stochastic simulation of seasonal intermittent streamflows for arid regions, Journal of Hydro-environment Research (2014), http://dx.doi.org/10.1016/j.jher.2014.06.001

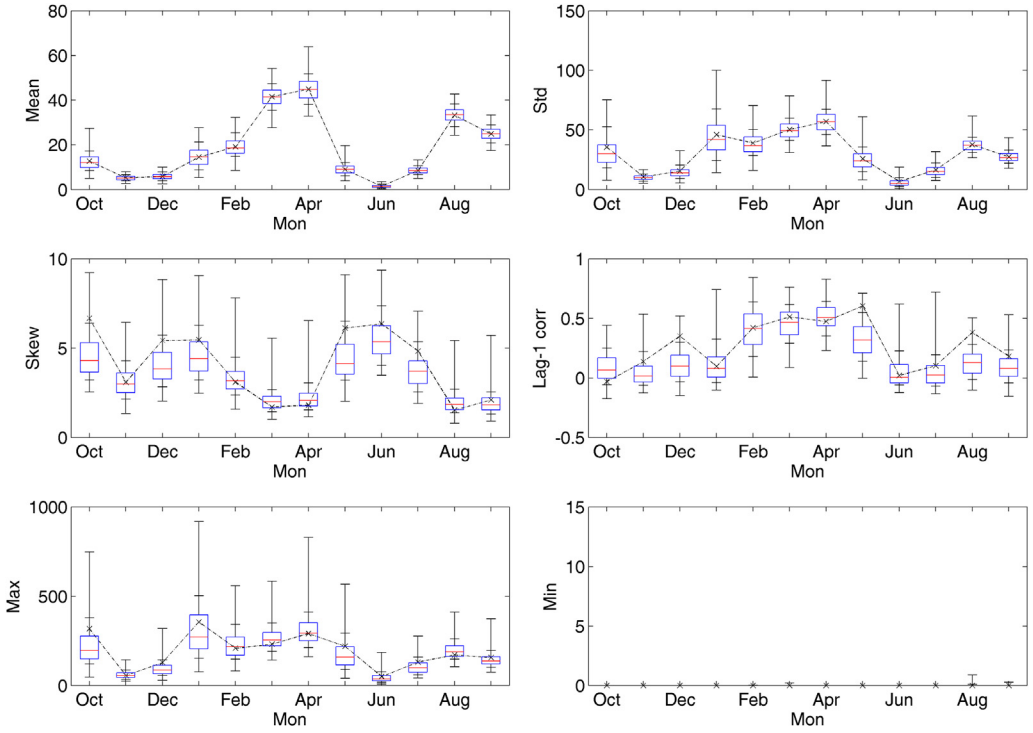

Fig. 2. Box plots of basic statistics for Colorado River monthly flows (million cubic meters, 10^6 m^3) obtained from data simulated using the BNCG model and from historical data (dotted line).

the minimum flow. Significant overestimations of the historical minima were observed among the data simulated using the PGAR model. The BNCG model better reproduced this extreme statistic. The densities estimated using Eq. (32) are shown in Fig. 8. The negative values of the estimated densities were induced during the smoothing procedure of Eq. (32). Note that

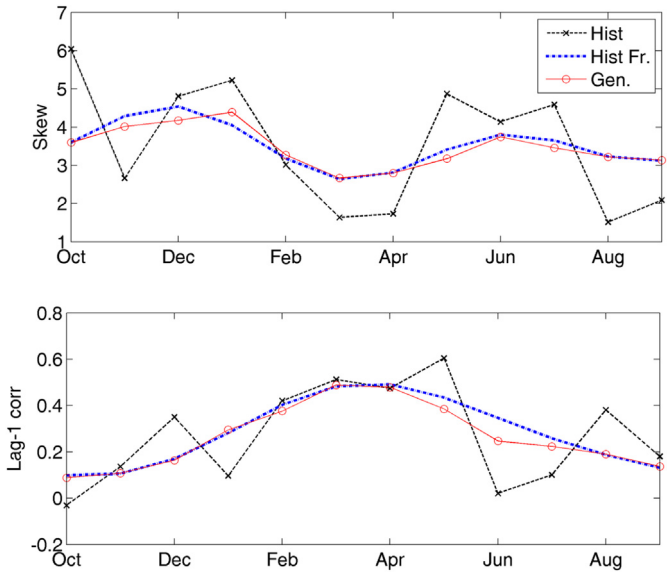

Fig. 3. Means of the skewness and lag-1 correlation obtained from the nonzero data simulated using the PGAR model (solid line with circles), from the historical data (dotted line with crosses), and the Fourier transformed statistics (thick dotted line).

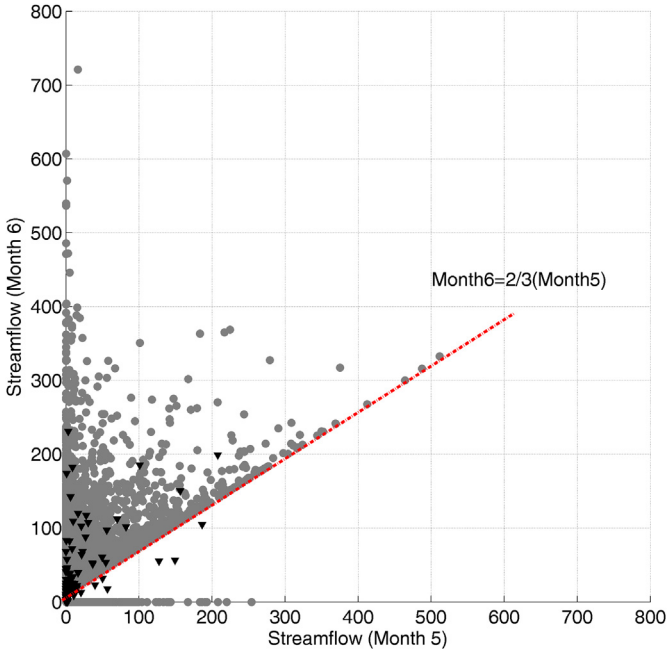

Fig. 4. Scatter plot of streamflows (unit: 10^6 m^3) of the Colorado River for month 5 (horizontal) and month 6 (vertical) derived from data simulated using the PGAR model (gray circles) and historical data (inverted triangles). The straight line splits the region where simulated values are not located in the lower region of the line except for the zero values, which are simulated from the occurrence process (X). The slope of the line corresponds to the lag-1 month-to-month correlation for month 6 (2/3 ≈ 0.67).

Please cite this article in press as: Jeong, C., Lee, T., Copula-based modeling and stochastic simulation of seasonal intermittent streamflows for arid regions, Journal of Hydro-environment Research (2014), http://dx.doi.org/10.1016/j.jher.2014.06.001

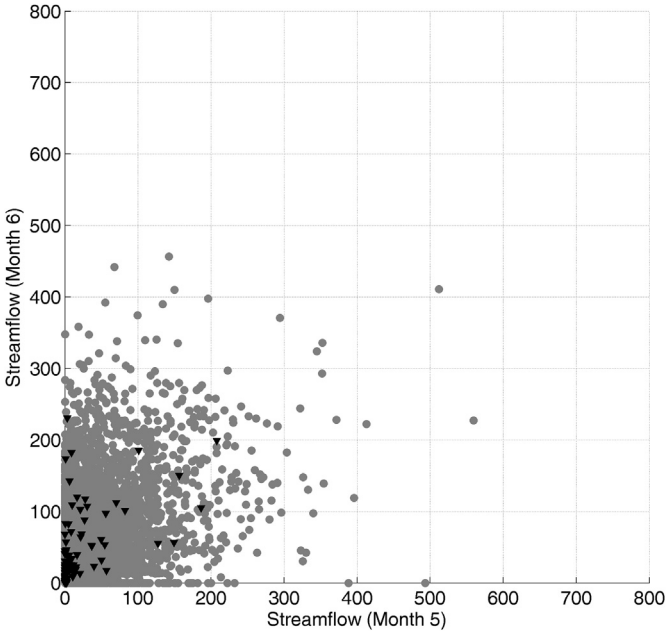

Fig. 5. Scatter plot of streamflows (unit: 10^6 m^3) of the Colorado River for month 5 (horizontal) and month 6 (vertical) derived from data simulated using the BNCG model (gray circles) and historical data (inverted triangles).

the PGAR model significantly underestimated the frequency of the lowest value. This underestimation is related to the underestimation of the minimum by the PGAR model noted above and implies that the data simulated using the PGAR model underestimated short-term (one or two-year) drought events.

Operational statistics, such as maximum consecutive length and amount for deficit and surplus, as well as storage capacity, were estimated using the PGAR and BNCG models (data not shown). All the operational statistics were accurately reproduced using the PGAR and BNCG models, excluding the maximum consecutive volume of deficit, which was underestimated by both models.

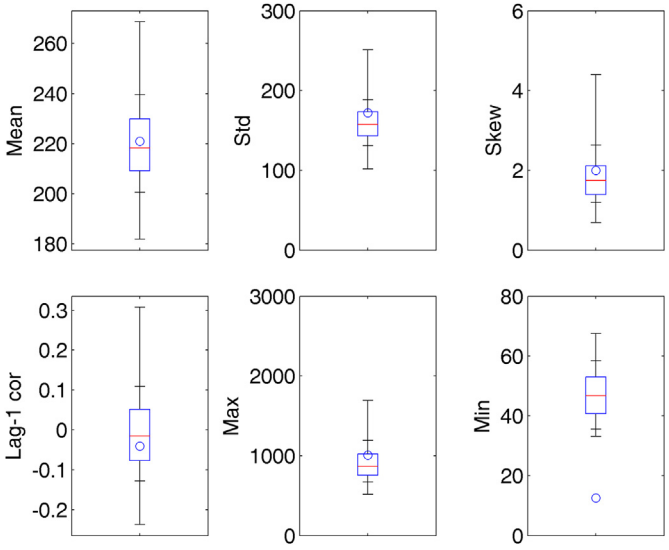

Fig. 6. Yearly key statistics obtained from flows simulated using the PGAR model and historical flow data (circles).

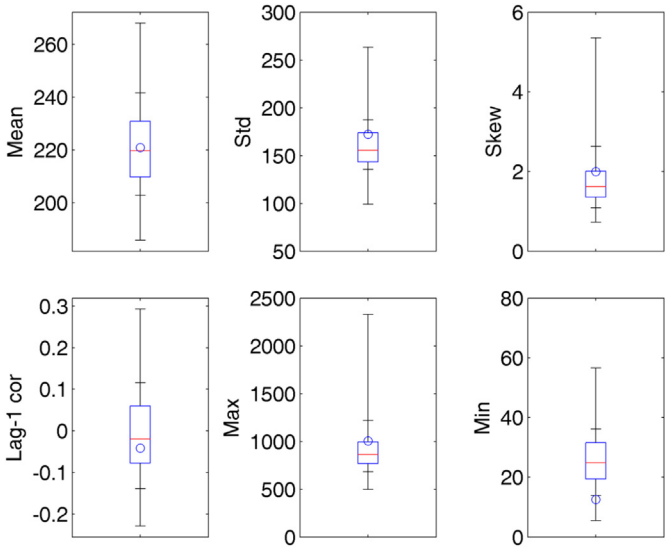

Fig. 7. Yearly key statistics obtained from flows simulated using the BNCG model and historical flow data (circles).

6. Results of alternative copula model

The main advantage of copula modeling is the flexibility of the relationship. Therefore, we further applied copula models to the monthly streamflow data of the Colorado River instead of BNC while the marginal distribution was fixed as a gamma distribution. The four copula models shown in Table 1 were considered. Different copula models were applied at each month by selecting from the highest maximum likelihood (equivalently lowest negative log-likelihood), denoted as the alternating copula model (ACM).

In Table 2, the estimated negative log-likelihoods are shown for each considered copula model. Note that the bolded values indicate the smallest value among the four copula

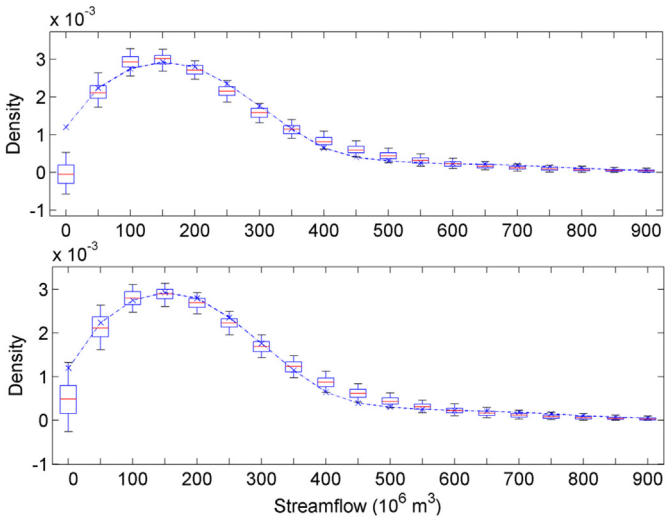

Fig. 8. Density estimates of yearly data obtained from simulated yearly flows (boxplots) based on the PGAR (top panel) and BNCG (bottom panel) models, as well as the historical data (dotted line with x marker).

Please cite this article in press as: Jeong, C., Lee, T., Copula-based modeling and stochastic simulation of seasonal intermittent streamflows for arid regions, Journal of Hydro-environment Research (2014), http://dx.doi.org/10.1016/j.jher.2014.06.001

Table 2
Negative log-likelihood from different copula models.

Month	Clayton	Frank	Gumbel	Gaussian
1	0.00[a]	−0.25	0.00[a]	**−0.28**
2	−0.13	−0.30	0.00[a]	**−0.40**
3	**−2.91**	−2.16	−2.53	−2.36
4	**−24.38**	−16.50	−7.07	−10.45
5	**−22.54**	−17.29	−11.08	−16.44
6	−11.77	**−20.38**	−16.72	−19.40
7	−16.27	−16.62	−15.55	**−17.79**
8	−11.61	−10.97	−12.13	**−14.32**
9	**−2.60**	−1.17	0.00[a]	−0.78
10	−1.50	−1.90	−0.88	**−2.30**
11	−6.81	−4.92	−5.39	**−6.92**
12	−0.18	−0.85	**−1.43**	−0.66

[a] Note that (1) bold type indicates the selected copula model with the smallest negative log-likelihood with the estimated parameters from IFM and (2) the values of 1.0 and 0.0 represent the estimated parameters that were not converged.

models. The copula parameters shown in Table 1 are presented in Table 3. The parameters were estimated with inference functions of margins (IFM). Lee and Salas (2011) can be referred to for further information related to IFM.

As an example, scatter plot and histogram of the CDF transformed historical data for Months 5 and 6 are illustrated in Fig. 9. Furthermore, the scatter plot of simulated streamflow between Month 5 and Month 6 is presented in Fig. 10. Note that Frank copula was selected for this month as shown in Table 2. Fig. 10 is comparable to Fig. 5 (i.e., the BNC model). The shape of the scatter with Frank copula (Fig. 10) is similar to BNC (Fig. 5). The behavior of the simulated key statistics is almost identical to BNCG because the marginal distribution is fixed as gamma (results not shown). The yearly operational statistics of the ACM are tested (data not shown). Compared to BNCG, the performance of ACM is similar.

Table 3
Estimated parameters from inference functions for margins (IFM).

Month	Clayton	Frank	Gumbel	Gaussian
1	0.00[a]	−0.56	1.00[a]	−0.08
2	0.16	0.61	1.00[a]	0.11
3	0.64	1.59	1.15	0.27
4	2.23	5.34	1.47	0.50
5	1.69	4.57	1.46	0.56
6	0.78	4.79	1.59	0.60
7	0.95	4.10	1.58	0.58
8	1.11	3.94	1.57	0.66
9	1.02	2.09	1.00[†]	0.26
10	0.46	2.03	1.15	0.32
11	0.63	2.08	1.26	0.39
12	0.06	0.83	1.12	0.12

[a] Note that the values of 1.0 and 0.0 represent the estimated parameters that were not converged.

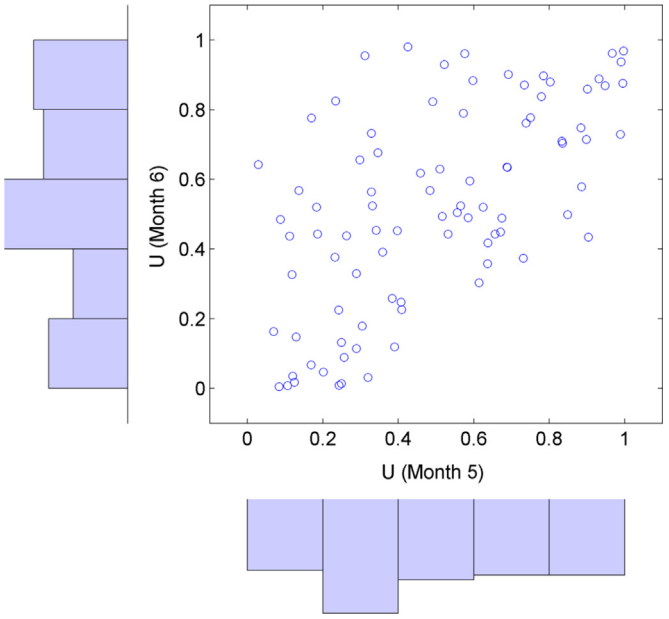

Fig. 9. Scatter plot and histograms of gamma CDF transformed data (U) for month 7 and month 8.

7. Summary and conclusions

Parametric stochastic models were tested to simulate intermittent monthly streamflows for a semi-arid or arid region. To reproduce intermittency in the simulated data, a product model of the amount and occurrence processes was applied. The PMC model was employed for the occurrence process, and two alternative models were tested for the amount process, the PGAR and BNCG models. Key statistics were estimated to validate the performance of the models over monthly and yearly time scales. Furthermore, ACM was tested to reveal the possible usage of alternating copulas.

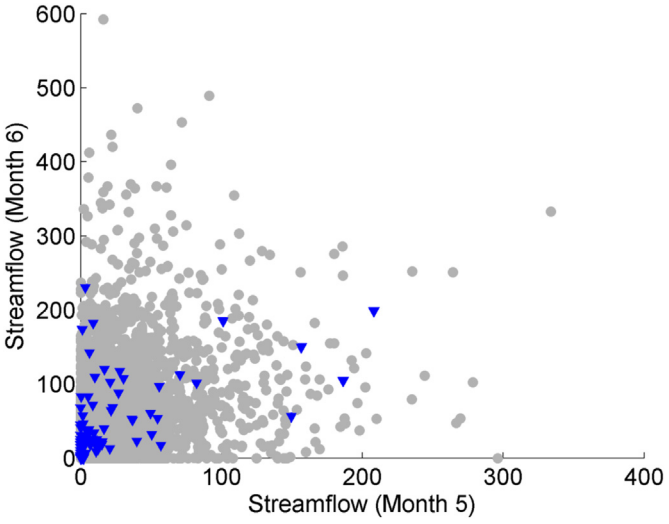

Fig. 10. Scatter plot of streamflows (unit: 10^6 m^3) of the Colorado River for month 5 (horizontal) and month 6 (vertical) derived from data simulated using the alternative copula model (gray circles) and historical data (inverted triangles).

Please cite this article in press as: Jeong, C., Lee, T., Copula-based modeling and stochastic simulation of seasonal intermittent streamflows for arid regions, Journal of Hydro-environment Research (2014), http://dx.doi.org/10.1016/j.jher.2014.06.001

A few drawbacks of the PGAR model were observed during the current study: (1) The parameter space was very restrictive, and therefore Fourier transformed statistics must be used in parameter estimation; (2) The simulated values were always greater than the previous monthly values due to multiplying the lag-1 correlation; and (3) Extreme events (i.e., minimum streamflows) were significantly underestimated on the yearly time scale, which resulted in underestimates of the numbers of short-term drought events. These drawbacks are general characteristics of the PGAR model.

In contrast, the copula model does not suffer from such drawbacks, as shown in the presented results. The marginal key statistics and operational statistics are well reproduced from the copula models. Additionally, the time-lagged relationship can be flexibly modeled following the relational shape of the historical data, as shown from ACM.

We conclude that the combination of the copula model of the amount process with the PMC model of the occurrence process is a reasonable approach for simulating intermittent monthly time series in semi-arid and arid regions.

Acknowledgment

We acknowledge that this work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2013-0362).

References

Chebaane, M., Salas, J.D., Boes, D.C., 1995. Product periodic autoregressive processes for modeling intermittent monthly streamflows. Water Resour. Res. 31, 1513−1518.

Chen, X.H., Fan, Y.Q., 2006. Estimation of copula-based semiparametric time series models. J. Econom. 130, 307−335.

Favre, A.C., El Adlouni, S., Perreault, L., Thiemonge, N., Bobee, B., 2004. Multivariate hydrological frequency analysis using copulas. Water Resour. Res. 40.

Fernandez, B., Salas, J.D., 1986. Periodic gamma autoregressive processes for operational hydrology. Water Resour. Res. 22, 1385−1396.

Joe, H., 1997. Multivariate models and multivariate dependence concepts. In: Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, London.

Kao, S.C., Govindaraju, R.S., 2008. Trivariate statistical analysis of extreme rainfall events via the Plackett family of copulas. Water Resour. Res. 44, W02415.

Katz, R.W., Parlange, M.B., 1995. Generalizations of chain-dependent processes − application to hourly precipitation. Water Resour. Res. 31, 1331−1341.

Kottegoda, N.T., Rosso, R., 2008. Applied Statistics for Civil and Environmental Engineers. Wiley-Blackwell, West Sussex, United Kingdom.

Koutsoyiannis, D., 1994. A stochastic disaggregation method for design storm and flood synthesis. J. Hydrol. 156, 193−225.

Koutsoyiannis, D., 2006. An entropic-stochastic representation of rainfall intermittency: the origin of clustering and persistence. Water Resour. Res. 42, W01401.

Kwon, H.H., Lall, U., Khalil, A.F., 2007. Stochastic simulation model for nonstationary time series using an autoregressive wavelet decomposition: applications to rainfall and temperature. Water Resour. Res. 43. W05407.

Lall, U., 1995. Recent advance in nonparametric function estimation − hydrologic application. Rev. Geophys. 33, 1093−1102.

Lall, U., Rajagopalan, B., Tarboton, D.G., 1996. A nonparametric wet/dry spell model for resampling daily precipitation. Water Resour. Res. 32, 2803−2823.

Lee, T., Ouarda, T.B.M.J., 2010. Long-term prediction of precipitation and hydrologic extremes with nonstationary oscillation processes. J. Geophys. Res. − Atmos. 115.

Lee, T., Salas, J.D., 2006. Record Extension of Monthly Flows for the Colorado River System. Bureau of Reclamation, U.S. Department of Interior.

Lee, T., Salas, J.D., 2011. Copula-based stochastic simulation of hydrological data applied to Nile River flows. Hydrol. Res. 42, 318−330.

Lee, T., Salas, J.D., Prairie, J., 2010. An enhanced nonparametric streamflow disaggregation model with genetic algorithm. Water Resour. Res. 46, W08545.

Loucks, D.P., Stedinger, J.R., Haith, D.A., 1981. Water Resources Systems Planning and Analysis. Prentice-Hall, Englewood Cliffs, NJ.

Nelsen, R.B., 1999. An Introduction to Copulas. Springer-Verlag, New York.

Ouarda, T.B.M.J., Labadie, J.W., Fontane, D.G., 1997. Indexed sequential hydrologic modeling for hydropower capacity estimation. J. Am. Water Resour. Assoc. 33, 1337−1349.

Salas, J.D., 1993. Analysis and modeling of hydrologic time series. In: Maidment, D.R. (Ed.), Handbook of Hydrology. McGraw-Hill, New York, pp. 19.11−19.72.

Salas, J.D., Abdelmohsen, M.W., 1993. Initialization for generating single-site and multisite low-order periodic autoregressive and moving average processes. Water Resour. Res. 29, 1771−1776.

Salas, J.D., Boes, D.C., 1980. Shifting level modeling of hydrologic series. Adv. Water Resour. 3, 59−63.

Salas, J.D., Lee, T., 2010. Nonparametric simulation of single-site seasonal streamflows. J. Hydrol. Eng. 15, 284−296.

Salas, J.D., Smith, R.A., Tabios, G.Q., Heo, J.-H., 2008. Statistical Computing Techniques in Water Resources and Environmental Engineering (Lecture Note). Department of Civil and Env. Engr., Colorado State University, Fort Collins.

Salas, J.D., Sveinsson, O.G., Lane, W.L., Frevert, D.K., 2006. Stochastic streamflow simulation using SAMS-2003. J. Irrig. Drain. Eng. − ASCE 132, 112−122.

Sharma, O.P., Le Treut, H., Seze, G., Fairhead, L., Sadourny, R., 1998. Interannual variations of summer monsoons: sensitivity to cloud radiative forcing. J. Clim. 11, 1883−1905.

Shiau, J.T., Feng, S., Nadaraiah, S., 2007. Assessment of hydrological droughts for the Yellow River, China, using copulas. Hydrol. Process. 21, 2157−2163.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer-Verlag, New York.

Song, S.B., Singh, V.P., 2010. Frequency analysis of droughts using the Plackett copula and parameter estimation by genetic algorithm. Stoch. Environ. Res. Risk A 24, 783−805.

Srinivas, V.V., Srinivasan, K., 2001. A hybrid stochastic model for multiseason streamflow simulation. Water Resour. Res. 37, 2537−2549.

Srinivas, V.V., Srinivasan, K., 2006. Hybrid matched-block bootstrap for stochastic simulation of multiseason streamflows. J. Hydrol. 329, 1−15.

Stedinger, J.R., Potter, K., Kibler, D.F., Tasker, G., 1983. Regional flood frequency estimation and network design − comment. Water Resour. Res. 19, 1343−1345.

Stedinger, J.R., Tasker, G.D., 1985. Regional hydrologic analysis. 1. Ordinary, weighted, and generalized least-squares compared. Water Resour. Res. 21, 1421−1432.

Stedinger, J.R., Vogel, R.M., 1984. Disaggregation procedures for generating serially correlated flow vectors. Water Resour. Res. 20, 47−56.

Sveinsson, O.G.B., Salas, J.D., Boes, D.C., Pielke, R.A., 2003. Modeling the dynamics of long-term variability of hydroclimatic processes. J. Hydrometeorol. 4, 489−505.

Wong, G., Lambert, M.F., Leonard, M., Metcalfe, A.V., 2010. Drought analysis using trivariate copulas conditional on climatic states. J. Hydrol. Eng. 15, 129−141.

Young, E.F., Holt, J.T., 2007. Prediction and analysis of long-term variability of temperature and salinity in the Irish Sea. J. Geophys. Res. − Oceans 112.

Zhang, L., Singh, V.P., 2006. Bivariate flood frequency analysis using the copula method. J. Hydrol. Eng. 11, 150−164.

Please cite this article in press as: Jeong, C., Lee, T., Copula-based modeling and stochastic simulation of seasonal intermittent streamflows for arid regions, Journal of Hydro-environment Research (2014), http://dx.doi.org/10.1016/j.jher.2014.06.001